December 2003

Problems in Designing Huge Datawarehouses and Datamarts

Didier Nakache
Cramif

PROBLEMS IN DESIGNING HUGE DATAWAREHOUSES AND DATAMARTS

Didier Nakache
Cramif
datamining@wanadoo.fr

Abstract

This paper reports on a Datawarehouse application. The French national health department has to face numerous problems: financial, medical, social, accounting, public health and political. Thus, an efficient tool is needed for managing the decision support information system. In this context we have proposed the ERASME/SNIIR-AM Datawarehouse project. To the best of our knowledge, it has been considered as the largest Datawarehouse in the world. The main challenge we had to solve were due to huge volumes. We have chosen to solve it by an ad hoc methodology mainly based on datamart design.

Keywords: Database, datawarehouse, decision-taking, datamart, datamining, KDD, health service, health

Introduction

The French National Health Service is responsible for a considerable amount of information and has to face many problems eg: availability and quality of data, heterogeneous sources, frequent updates for a large amount of information, different ways of computing the same information according to different points of view, etc. Moreover the political context and rules mean that Health Service needs the latest tools to analyze data and send information to it's partners. Lastly, economic context means that the institution must improve it's spending to achieve a minimum of break-even. Taking into account these elements, the ERASME datawarehouse project was designed. To the best of our knowledge, it has been considered as the "largest datawarehouse in the world".

The purpose of this paper is to report on this experience. Section 2 presents the context and the objectives of the project. Section 3 exposes different problems with health information: functional, and design. Section 4 presents costs and technical data of a prototype. Section 5 presents the general architecture. Section 6 is devoted to the design methodology, and finally, section 7 presents some results of the prototype.

The Context and the Objectives

The general regime covers all salaried workers, about 80% of the population and represents:

- 100 000 health service employees,
- 47 million "clients,"
- 1 billion invoices per year,
- 100 billion dollars in annual turnover.

The ERASME/SNIIRAM project covers all French social security regimes. In other words the entire population (58 millions) is concerned.

The system has many objectives at local, regional and national level: to carry-out operations and analyses under the scope of cost and internal control, perform researches and analyzes to improve spending awareness as well as the application of sanitary studies,

to publish official information, adapted to each category of recipient… in other words, the main objective is to manage efficiently the health system by its different partners (government, health service, ...).

The Problems

Problems are numerous but can be summed up in one sentence: how can the Health Service be improved? This question covers several aspects:

- Accounts: How can we be sure that Health Service spending is efficiently monitored?
- Political: How can we legislate? What costs would be incurred by introducing new measures? How can we supply opposable and shareable data to partners?
- Financial: how can we improve healthcare at less cost?
- Public health: do we have good healthcare?
- To understand more clearly what's at stake, a 1% error represents 1 billion dollars.

Example of Problems

The problems treated by the Health Service are particularly complex. Let's consider the example of pathology and medicine and try to reply to a simple question: how can we treat 'X' (e.g.: the flu) in France in 2003? First is the problem of coding: is the illness identified? Is it possible to reach the diagnosis using a different code? Are the acts coded in such a way that they can be analyzed? Does the corresponding governmental body authorize a statistical treatment?

Even if we could manage to answer these questions, we would then be confronted to a second problem: Health Service is mainly interested in what is covered and what is reimbursed (purely financial) rather than what is really spend for health. Many players (chemists, social security agents, practitioners, health insurance contributors) do not believe it is useful to record information where reimbursements are not involved. A Sickness Note covering less than three days is not reimbursed (obligatory waiting period). The Social Security is not always notified if an individual comes to an arrangement with his company.

Turning now to the question of what the chemist provides: there is often a big difference between what is written on the prescription and what the chemist actually gives the patient as he is authorized to substitute one medicine with another if it is cheaper (generic drug). This means that for two different prescriptions the chemist can hand out exactly the same medicine. If these two obstacles are removed, how can we compare two prescriptions? About 15,000 medicines exist in France: many of which appear in different forms, have many variants, the packaging is different or they are not administered in the same way. Classification systems exist (ATC, EPHMRA, etc.) but the drugs are generally filled by either their chemical or anatomic principal active ingredient, and for example none of these helps to directly identify the antibiotic!

It should also be added that no one way exists to identify an individual. In France, only workers are knowed, and other people are covered by them. If we take the example of a young woman who is either at her parent's charge or a student: she is entitled to healthcare cover as her parents child or in her own right as a student. She gets married and may become her husband's responsibility therefore changing her number and also her social security regime if she was a student. She later finds work and finally obtains her own number in her own right. However, if she is made redundant she can reverts back to being at her husband's charge. Some precarious situations mean many changes.

Doctors are identified by one number made up of the department and a number attributed to them by the Medical Practice Council: a doctor who changes department changes his number!

How can we analyze the request of someone who submits his claim for reimbursement for the chemist one year later when the laboratory tests were reimbursed very early on (direct payment for example)?

Finally, how can we address the problem of errors (mostly typing), which lead to extra/duplicate reimbursements: the accountant would like to manage the cash flow; the analyst would like to manage the cost.

How can we evaluate the quality of treatment? The economic factor is certainly important but is it sufficient? Other criteria like secondary effects, the length of the illness (with or without work stoppage), the risk of it happening again, the degree of suffering … all these are equally important factors difficult to quantify.

Technical and Design Problems

If we had to design datawarehouse from the requirements, each user should have tell he needs data and information about practitioners, spends, reimbursements, … but a traditional analysis would have generate too large volumes and only one big datawarehouse. If we wish to exploit the data directly from the datawarehouse (which contains all the detailed information), any request would ask for dozens hours of process. Every month, about 400 heavy requests have to run, to which you have to add studies and punctual analyses. The time for answer would not be compatible with needs and the processing of one month would need more than a month of delay. It was thus necessary to find a solution which allows to cover the requirements by by-passing problem of volume. Information about eighteen to twenty-four months represents about 100 terabytes. When the project was initiated by the Health Minister in 1997 no information system could store such a volume. So an original design methodology has been used to create datamarts with acceptable volume of data.

Technical Information

The Prototype

When the architecture was defined no computer system had the capacity to store the information therefore a prototype was deemed necessary in order to be able to validate the technical choices based on the following configuration: a SUN E 10 000 computer with 18 processors at 336 Mhz, with 12 gigabytes of RAM and 2,5 terabytes of disk space (386 disks of 9 Go - RAID 5 technology). For the software, Oracle 8i and UNIX as the operating system was installed. The prototype acted as a benchmark (3 months) to choose the tools for extraction, loading, requesting information, datamining and reporting.

The Cost

The global cost of the project is 43 million dollars (human cost not included) for an estimated workload of about 200 man years. The total estimated return on investment 5 years after the beginning is about 750 million dollars.

General Architecture of the Application

The database is centralized with only one interface supplying the information considered as relevant for the datawarehouse. Basic information is gathered from the computer centers, local insurance companies and other parts of the health service. Datas are received every day and controls are carried out at a higher level and done before payment wherever possible. They are included in the national datawarehouse and datamarts and duplicated at regional level. Each datawarehouse contains elementary information and is not generally for use, the only exception being to send data to the datamarts which themselves contain official and detailed information. Regional datawarehouse has the same structure as national but with less data (only the geographical region).

Design Methodology

When the project began (in 1997), there was no design methodology specialized in datawarehouse design which had penetrated the industrial sector in France, but it has been and still is subject of various research projects ([Inmon 96], [Kimball 1997], [Kimball 1998], [Cabibbo and Torlone 1998], [Golfarelli and Rizzi 1998], [Golfarelli 1999], [Moodu and Kortnk 2000] [Giovinazzo 2000], [Akoka et al. 2001]). An interesting state of the art can be found in [Laender 2002]).

The ERASME datawarehouse was built according to a methodology very close to X-META [Carneiro 2002] which proposes an implementation by successive iterations, according to principles of Boehm's spiral (management of the risk) associated with RAD techniques, in particular: prototyping, architecture, metadata, tools, pilot project, then datamarts and datawarehouse. This method is interesting in the field of the project control, but lack of methodology for an optimization of the datamart's contains in order to increase performances.

Figure 1. General Architecture of the 13 Datawarehouses

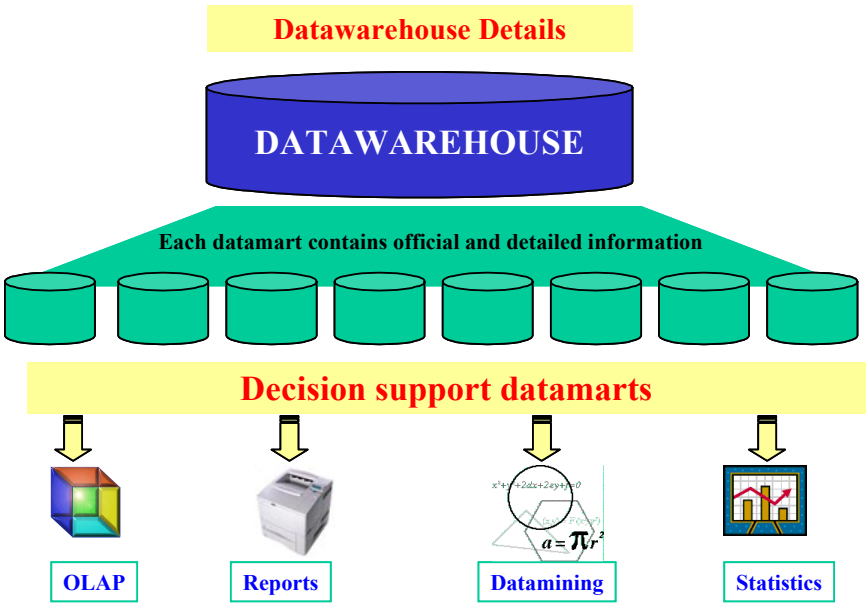

Figure 2. Details of One Datawarehouse

Related Works

Some global approaches for datawarehouse design have been proposed. Kimball [Kimball 1998] [Kimball 1997], describes a complete framework for applying dimensional modeling, called Business Dimensional Lifecycle. The author suggest interviews and facilitated sessions as the basic techniques for this step. The difficulty of identifying dimensions is also addressed by Kimball in [Kimball 2000]. Pereira [Pereira 2000] presents a global methodology for development of Datawarehouse pilot projects, successfully tested in a military datawarehouse project.

Some approaches are more traditional: Senman proposes a design process similar to traditional database design process [Senman 2000]. Agrawal and Cabibbo [Agrawal 1997] [Cabibbo 1998] think that at least a conceptual or logical model should precede the implementation of a datawarehouse. Lambert [Lambert 1995] proposes to get the user more involved. The project and method X-META [Carneiro 2002] is successfully based on this idea.

A promising approach for multidimensional design consists in starting from OO, UML or ER conceptual schemas. Krippendorf and Song [Krippendorf 1997] address the problem of mapping an ER schema into a star schema, and Song et al. [Song 2001] analyze many-to-many relationships between fact and dimension tables. An extension of the ER model for capturing the dimensional paradigm is presented in [Sapia 1998]. Moody [Moody 2000] explored the techniques for deriving the multidimensional models from traditional entity relationship schemas and relational data bases. Akoka [Akoka et Al. 2001] proposes an approach, which consists in designing datawarehouse, dimension hierarchies and datamarts from UML schemes. The dimensional fact model, proposed by Golfarelli et al. [Golfarelli 1998], provides a graphical notation for representing datawarehouses based on the dimensional model. The mapping of ER schemas into this model is discussed in [Golfarelli et Al.1998]. In [Golfarelli 1999], he describes the design of datawarehouses according to the dimensional fact model. Phipps and Davis [Phipps 2002] propose a method to design a datawarehouse starting from an entity relationship model in five steps: (i) Find entities with numeric fields and create a fact node for each identified entity, create numeric attributes of each fact node, (ii) based on numeric fields in the entities, (iii) create date and or time levels (dimensions) with any date/time type fields per fact node, (iv) create a level (dimension) containing the remaining entity attributes (non-numeric, non-key, and non date fields), and finally (v) recursively examine the relationships of the entities to add additional levels in a hierarchical manner. Some other methodologies approaches are object-oriented based [Firestone 1998, Giovinazzo 2000, Trujillo 2001]. Many object oriented approaches emphasize the user involvement in early stages of requirements analyzis. These approaches don't take into account sufficiently the problems of optimization of important volumes.

Few researchers focused on complex problems and optimization of volumes. Rizzi more specifically considered the optimization of volumes and indexes in a constrained space [Rizzi 2002]. To this end, he created a model based on a calculation of costs to optimize the creation of index for a given global space constraint. Rokia Missaoui [Missaoui 2000] focuses on recent developments in representational and processing aspects of complex data-intensive applications.

Others are proposing tools. Laender [Laender 2002] created a tool (MD-2) based on the dimensional data modeling approach, which facilitates the user participation in the development of a datawarehouse application. MD2 assists users in identifying their analytical requirements in order to help datawarehouse designers to better specify business requirements and translate them into appropriate design elements. This approach is strongly centered on the metadata repository used to build the dimensional schema and to specify reports that can be generated from the datawarehouse and visualized through a Web interface.

Toward a Methodology Devoted to Huge Volumes

The initial problem was the following one: 1 billion invoices a year represent approximately 5 to 10 billion records a year. They correspond to prescriptions made by about 250.000 practitioners for 60 million potential patients, allowing them to care 50.000 diseases. By multiplying these data by the number of years of history (to be in accordance with the French legislation which allows to be paid off during two years and three months), it represents a very important volume of data.

If we wish to exploit the data directly from the datawarehouse (which contains all the detailed information), any request would require dozens hours of process. Every month, about 400 heavy requests have to run, to which must be added the studies and punctual analyses. Response time would not be compatible with needs and the processing of one month would need more than one month of delay. It was thus necessary to find a solution which allows to cover the requirements by by-passing problem of volume.

Remark 1: A traditional approach by trade datamart would have given no satisfactory result because 80 % of requirements use common data (access to the data of benefits, practitioners, establishments, and patients).

Remark 2: The snowflake model is used as an alternative to the star schema for dealing with large amount of data. However, in our application, problems of volume are located on the facts table rather than on the dimension tables. So this can't be a solution to our problem.

Thus we have proposed our own methodology. Our approach consists in designing several datamarts with common axes but different optimized aggregations. This technique also favored the management of the rights and the security. For example, the data relative to the pathologies are very sensitive and confidential while the data for geographic sanitary card are accessible to all, including persons outside the institution. The question then became: how to organize in best aggregated data and detailed data to answer all these imperatives (functional requirements, security, volumes and response time). The solution consisted in breaking the problem in roughly fragments (first level datamarts) and in breaking each fragment in smaller fragments (second level datamarts). The method considers the crossings of axes; it consists in determining for each request the entities (or multidimensional axes) concerned, without worrying about data (attributes). Main axes extracted from the analysis are: practitioners, benefits, beneficiaries, establishments and structures, pathologies… To determine the first level datamarts, it was necessary to identify the crossings of axes with the strongest functional weight. We determined, for example, the following datamarts of first level: offer of care, consumption of care, follow up of countable expenditure, analyze of pathologies, … To determine the second level datamarts, we were interested in the possibilities of aggregation, rights, and access… But in a restricted domain, it is easy to compose with these parameters. Only one question remained unresolved, concerning view materialization: must datamarts (first and second level) be physically built or just views from the databases? This question has been taken under consideration case by case, for each datamart. Avantage of this solution is to propose users small datamarts, easy to use, with short answer times. The disadvantage of this method is to propose different datamarts for same users. So one user will have to reach several datamarts according to his trade and his requirements.

The Proposed Design Methodology

The process followed for the design can be summarized by these different steps (Figure 3):

Strategic objectives ➪operational objectives ➪ requests ➪ axes and attributes ➪ crosses of axes ➪ functional weight ➪ first level of datamarts ➪ second level of datamarts.

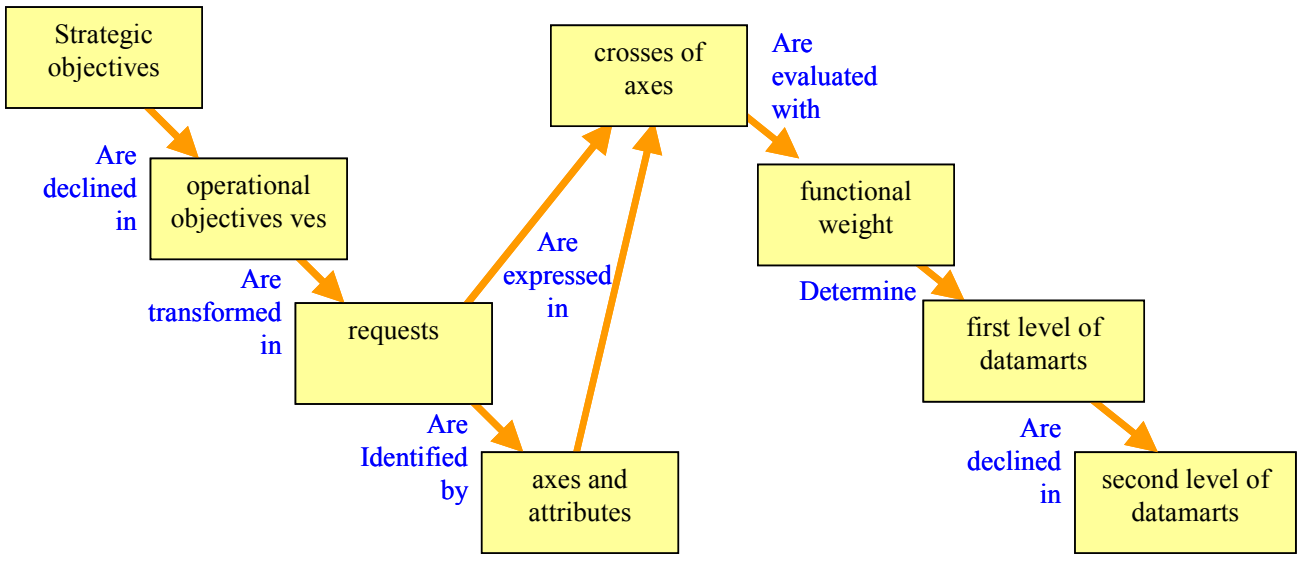

Figure 3. Steps of Design Methodology

Some Definitions

Strategic objective: a variation of a fundamental mission. Example: Countable follow-up of the expenditure for health.

Operational objective: an operational variation of a strategic objective: either an essential mission, or the same requirements expressed by several sectors. Example: follow-up of the expenditure of health per countable account compared to the budget. The identification of an operational objective arises from the interviews.

Request: a request (do not to confuse with its expression in SQL for example) is the detailed variation of an operational objective, but with no detailed specifications. For example, the follow-up of the expenditure of health per countable account is expressed by several requests, of which: measure each month the amount of expenditure by countable account, with recall of the last twelve months and comparison with the budget.

Functional field: the logical regrouping of several operational objectives at the level of the company. The finality is to facilitate the research and the presentation of work. Examples: countable spend, consumption analyze, follow-up of pathologies. It is generated by a synthesis.

Axes of analysis: an axe of analysis is a dimension, but from a conceptual point of view. It is specified by attributes. They are obtained from the operational objectives. For each objective, the axes of analysis are defined. An empirical method to identify them consists in asking the following question "what do I want to know for this objective and how?". Each time we meet the word "BY" in an answer,we obtain either an attribute or an axe of analysis. So you can specify the useful attributes. Example: "I want to know the spends for health by patient and by practitioner to realize the monitoring tables of the expenditure of health". In this case "Patient" and "practitioner" become axes of analysis. (NB: we often find behind the word "FOR" whole or part of the objective). Examples of axes: patients, practitioner, organizations...

Attribute: an elementary variation of an axe. It is very close to the concept of physical field of a table, but the translation from the attribute into field is not systematic. The physical form of the data can be modified for reasons of volume, optimization, cardinality... At this stage, we are still at a conceptual and logical level. For example, for the "patient" axe, we have the following attributes: identification / date of birth / sex / Socio Professional Category...

Steps of the Method

Our method for datamart's design is articulated in three phases: collect the objectives and the requirements, express them into requests and axes, and define the datamarts by crossing axes. The essential word who prevailed with the design was: "Iteration". The analysis didn't start by the expression of requirements but rather by the analysis of various documents: the Health Service's task, its strategic plans, the law regarding social security (national health service) financing, the contracts regarding objectives and management…

Stage 1: Collecting the objectives and the requirements

Collecting the requirements was held in three steps: the common base, the analysis of the strategic objectives by direction and the collecting of the desired products. A first document (common base) was filled out by each direction. Its finality is to position the trade to the project, the strategic objectives, the legislative environment, and to decline the strategic objectives in operational objectives within the framework of the project. The questions were: definition of the trade, essential missions, position of the trade with the strategic orientations of the health service, the legislative and lawful contexts, identification of the texts supporting or slowing down the action, identification of the lawful or legislative lacks…

The second phase is the analysis of the objectives. Each direction filled out a file by objective, and declined the requirements of knowledge, the type of awaited result and the actors. The objectives were classified by priority.

At least, requirements were declined in expected products. In this step, we specify data sources, granularity... Collected requirements have been reformulated, compiled and summarized. Reformulation made possible to highlight common requirements (or very close) in different sectors. Summary made appear 26 functional objectives, for example: Follow-up of the expenditure of health, demography of the offer of care, follow-up of the individual activity of practitioners, analyzes the consumption of medicines...

Stage 2: Requests and axes

Each objective is declined in requirements. From these needs, the standard requests and the potential axes of analysis are identified. It is from this step that the first pivot takes place: each need is declined in requests, attributes and axes. These last points become then the center of the analyze. It was necessary to create a document which specifies, for each request, axes (by specifying if they are aggregated or detailed), attributes, and objective or covered requirements. It is useful at this stage to use the previously defined functional regroupings. This analysis will allow us to redefine the functional perimeter of the datawarehouse, to create the axes, to create a meta model, to begin to build the metadata repository and in finally to constitute the datamarts. The meta model and the repository offer a good visibility of the system. All the necessary attributes are extracted from the analysis of the requests and constitute the repository, which is fed from the very start of the project.

Stage 3: Constituting the datamarts

Constituting datamarts is done in several steps:

- Identify crossings of axes: starting from the requests and for each one, identify axes to be crossed, without specifying any level of aggregation. Note the number of axes to be crossed as well as the importance of the functional requirements (strong, intermediary, secondary)

- Calculate the functional weight of each cross of axes: from the previous step, by multiplying the importance of requirements (affected with coefficients 20, 5 and 1 according to its importance) by the number of requests concerned. Then sort the results in descending order to treat in priority the strong and complex requirements.

- Create the first level datamarts: when the functional weight is computed, it is enough to make functional groups. The regroupings are made from the weights, the requirement, the analysis and the users (right and trades). You can constitute a datamart, even weak weight, if it is intended for a particular category. The regroupings represent the first level datamarts. Example (extract) of weight, crosses, and first level datamarts:

Table 1. Example of Result (Cross of Axes)

Axes	Secondary need	Intermediate need	Strong need	number of requests	number of axes	Weight	DATAMART
Patient X Establishment X Benefits X Practitioner		4	14	18	4	1200	Consum
Establishment x Benefits x Practitioner	1	9	17	27	3	1158	Offers
Benefits x Practitioner	2	6	22	30	2	944	Offers
Establishment x Benefits	1	9	17	27	2	772	Offers
Patient x Benefits	4	7	11	22	2	518	Consum
Patient x Benefits x Practitioner	1	2	7	10	3	453	Consum
Patient x Establishment x Benefits			7	7	3	420	Consum
Patient x Pathology x Benefits		2	3	5	3	210	Patho

Consum = consumption of care, Offers = offers care (the patient axis does not intervene), Patho = pathologies

- Constitute second level datamarts: when datamarts of first levels are designed, it is necessary to create datamarts of second level, by using analysis, requirements and users (or trades), but more especially granularity and needs of aggregation. It becomes easier because at this step, we are in a very restricted functional field.

- Consolidate: to validate results, obtained by successive iterations, requests are reallocated in the datamarts.

Some Results

Nevertheless some analyses have been carried out using the prototype. Here are some examples about medicine: a Kohonen card, a hierarchical ascending classification and a neural network analysis. These studies were based on reimbursements over two years using only the date of reimbursement and the medicine's code. These elements were joined to the medicines' file which contained other information (in particular the ATC and EPHMRA classifications).

This approach may seem simple but is not without interest. Certainly over the years the results have surprised the doctors who find them strongly redoubtable. Nevertheless on observing the Kohonen card it can be seen that on the lower part and a little on the right hand part, prescribed medicines have been strongly influenced by the substitution of generics.

NEURONE CLASSIFICATION (KOHONEN NETWORK)

CLONIDINE, CROMOGLICIQUE ACIDE DIHYDROERGOTAMINE, DIHYDROERGOTOXINE, DICLOFENAC, DILTIAZEM, FLUTAMIDE, DIPYRIDAMOLE, ISOPYRAMIDE, INDAPAMIDE, FUROSEMIDE, GLICAZIDE, ISOSORBIDE DINITRATE, METHYLDOPA, INDOMETACINE, NAFTIDROFURYL, NIFEDIPINE, PIROXICAM, RANITIDINE, PIRACETAM, SULPIRIDE, SELEGILINE, TIMOL, TRAMADOL, TRANEXAMIQUE ACIDE, TRIMETAZIDINE,	ACEBUTOLOL, AMIODARONE BROMOCRIPTINE, DIAZEPAM, MIANSERINE, NICERGOLINE, PROPRANOLOL, SPIRONOLACTONE, HALTIZIDE, TAMOXIFENE, VALPROIQUE ACIDE,	BACLOFENE, BUFLOMEDIL, DIOSMINE, ECONAZOLE (NITRATE), FLUNITRAZEPAM, KETOPROFENE, METOPROLOL (Tartrate de), SPIRONOLACTONE, VERAPAMIL,
ACICLOVIR, CIMETIDINE, CYCLANDELATE, PENTOXIFYLLINE, SUCRALFATE, SULFAMETHOXAZOLE, TRIMETHOPRIME,	ALLOPURINOL, ALPRAZOLAM, BROMAZEPAM, OXYBUTYNINE, ZOPICLONE,	ATENOLOL, BETAHISTINE, METFORMINE, TIAPRIDE,
AMBROXOL	AMOXICILLINE, ERYTHROMYCINE ETHYLSUCCINATE, ERYTHROMYCINE PROPIONATE, IBUPROFENE, LOPERAMIDE, (CHLORHYDRATE), PHLOROGLUCINOL,	BUSPIRONE, TRIMEBUTINE,	CALCITONINE, FENOFIBRATE, LACTULOSE, MEBEVERINE, THIOCOLCHICOSIDE,
CARBOCISTEINE, CEFADROXIL, CEFALEXINE, CEFRADINE, DOXYCYCLINE, NIFUROXAZIDE,		DEXTROPROPOXYPHENE, PYRANTEL, TETRACYCLINE,	MINOCYCLINE, TETRAZEPAM, TIAPROFENIQUE ACIDE,

Figure 5. Kohonen Card – Medicines

A Kohonen card concerning molecules and principal active ingredients can enable the detection of niches and could influence laboratory research.

The second graph is equally interesting: atypical behavior appears quite clearly for three categories of medicine (Dextropropoxyphene, Amoxicilline, Carbocistéine and very slightly for Buflomedil). It seems that during this period their reimbursement was modified (non-reimbursable or reduced from 65% to 35%) or they were criticized in the press for « being almost ineffective » or replaced by other generic medicines.

An analysis into the way a certain medicine was taken some years ago showed an atypical behavior as to when it was prescribed. The medicine concerned was particularly taken in spring, mostly by women. A medical enquiry showed that the medicine had diuretic and slimming properties (even though it wasn't prescribed for these reasons) and, with the approach of summer, many people had it prescribed to help them lose weight.

Certain questions however don't have answers. Take for example the study done several years ago which showed that when a surgeon settled in a region which hadn't previously had a surgeon, the number of operations rose considerably. What conclusion should be drawn ? Was the surgeon someone who created his « clientele (patients) » or did the very presence of a surgeon save lives, avoiding suffering and complications?

According to the experts if numerous studies are carried out the people doing them need to be supervised. Hospitals operate on a « global budget » principle which means that the budget has been attributed to them for the current financial exercise. For certain items where the budget is restrained and/or non-existent the hospital can prescribe them but the patient collects them in town. The most well-known example of this is x-rays. Only by supervising the patients was it possible to see if the x-ray was relevant and should it have been done in hospital ? This is what the Health Service accountants call « transferring between envelopes ». The detection and analysis of transfers causes many problems with statistics alone.

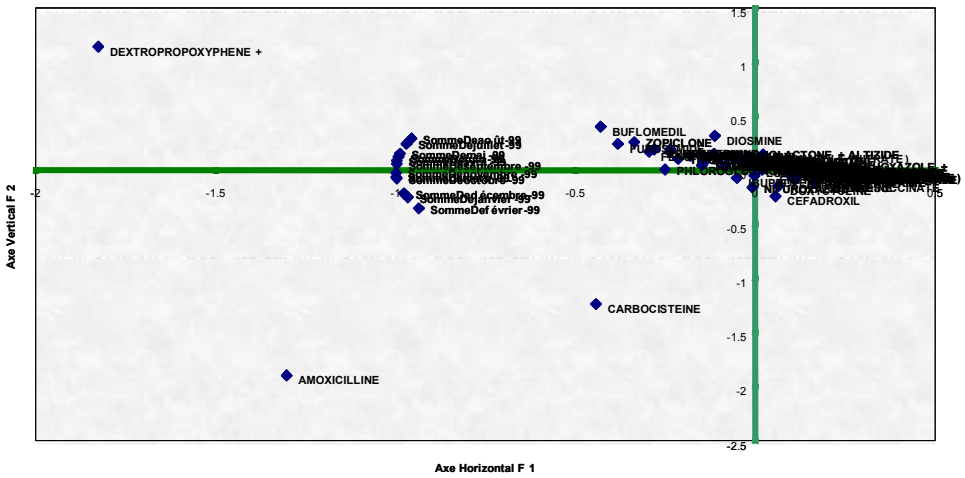

Figure 6. Analysis of the Principal Components of the Medicines

To end, here is one approach in trying to identify the difference between two prescriptions. The basic question is: how can we compare the two prescriptions?

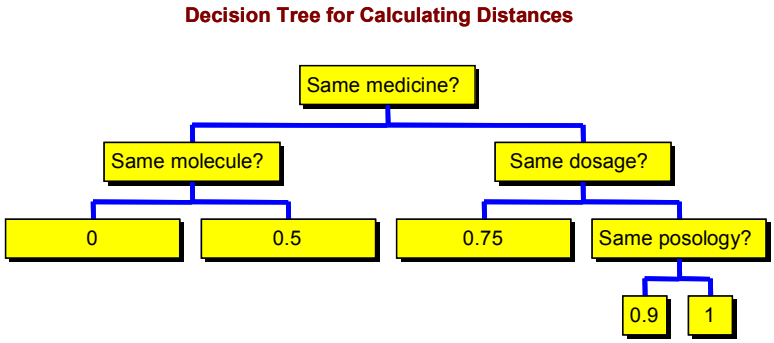

Figure 7. Calculating the Distances between the Prescriptions

Conclusion

The realization of this warehouse represents an important technological and political challenge. Putting it into practice is progressive using datamart and the first results must provide the elements essential in replying to multiple problems and lead us to the end result: how to treat illness at minimal cost.

Nevertheless, there are still several technical problems to solve: how can we effectively compare two prescriptions and, in particular, which guidelines should be established when considering two similar prescriptions ? Should the datamarts just be views of the datawarehouse or physical structure ? How is it possible to efficiently update the (huge) flow of patient/insured information, the reimbursements … How can we carry out long-term studies on sample databases (government body constraints) which will enable us to determine the patients' treatment, how do you define the sampling procedures which will provide sufficient information to meet the needs which may be expressed in 1, 5, 10, 20 years ? How also can we identify healthcare outbreaks; qualify them, arrange them in order, give them a signification in terms of treatment processes (preventative, curative, follow-up)? How can we make the information readable by outside users (non Health Service personnel) and transform the database from general information (accounting rectification, illegible nomenclatures) to statistics ? Finally, how can we optimize the matching up of individual, anonymous, external information?

References

[Agrawal 1997] R. Agrawal, A. Guppta, and S. Sarawagi: Modeling Multidimensional Databases. Proceedings of the Thirteenth International Conference on Data Engineering, Birmingham, UK, 1997, pp. 232-243.

[Akoka et al. 2001] J. Akoka, I. Comyn-Wattiau and N. Prat: "Dimension Hierarchies Design from UML Generalizations and Aggregations", ER'2001.

[Cabibbo 1998] L. Cabibbo, and R. Torlone: A Logical Approach to Multidimensional Databases. Proc. of the 6th Int'l Conference on Extended Database Technology (EDBT'98), Valencia, Spain, March 1998, pp. 187-197.

[Carneiro 2002] L. Carneiro, A. Brayner: X-META - A Methodology for Datawarehouse Design with Metadata Management – Caise 2002.

[Firestone 1998] J.M. Firestone: Dimensional Object Modeling. Executive Information Systems, Inc., White Paper n. 7, April 1998 (available at **http://www.dkms.com/ DOM.htm**).

[Gardner 1998] S. R. Gardner: Building the Datawarehouse. Communications of the ACM, v. 41, n. 9, p. 52-60. September, 1998.

[Giovinazzo 2000] W. A. Giovinazzo: Object-Oriented Datawarehouse Design. Prentice Hall, New Jersey, NJ, 2000.

[Golfarelli and Rizzi. 1998] M. Golfarelli, S. Rizzi: A methodological framework for data warehousing design, ACM workshop on data warehousing and OLAP, 1998.

[Golfarelli 1998] M. Golfarelli, D. Maio, and S. Rizzi: The dimensional fact model: a conceptual model for datawarehouses. International Journal of Cooperative Information Systems 7, 2-3 (1998), 215-247.

[Golfarelli et Al.1998] M. Golfarelli, D. Maio, and S. Rizzi: Conceptual Design of Datawarehouses from E/R Schemas. Proc. of the 31st Hawaii Int'l Conference on System Sciences, Vol. VII, Kona, Hawaii, 1998, pp. 334-343.

[Golfarelli 1999] M. Golfarelli, and S. Rizzi: Designing datawarehouses: key steps and crucial issues. Journal of Computer Science and Information Management 2, 3 (1999).

[Inmon 1996] W. H. Inmon "Building the Datawarehouse", John Wiley and Son editors, ISBN: 0471141615, 1996.

[Kimball 1997] R. Kimball: A Dimensional Modeling manifest. DBMS 10, 9 (August 1997).

[Kimball 1998] R. Kimball, L. Reeves, M. Ross, and W. Thomthwaite: The Datawarehouse Lifecycle Toolkit: Tools and Techniques for Designing, Developing and Deploying Datawarehouses. John Wiley & Sons, New York, 1998.

[Kimball 2000] R. Kimball: Mystery Dimensions. Intelligent Enterprise Magazine 3, 5 (March 2000).

[Krippendorf 1997] M. Krippendorf, and I-Y. Song: The Translation of Star Schema into Entity Relationship Diagrams. Proc. of the Eighth Int'l Workshop on Database and Expert Systems Applications, DEXA'97, Toulouse, France, 1997, pp. 390-395.

[Laender 2002] A. H. F. Laender, G. M. Freitas, M. L. Campos: MD2 – Getting Users Involved in the Development of Datawarehouse Applications – Caise 2002.

[Lambert 1995] B. Lambert: Break Old Habits To Define Data Warehousing Requirements. Data Management Review (December 1995).

[Missaoui 2000] R. Missaoui, R. Godin, J.M. Gagnon: Mapping an Extended Entity-Relationship into a Schema of Complex Objects. Advances in Object-Oriented Data Modeling 2000: 107-130.

[Moody 2000] L.D. Moody, and M.A.R. Kortink: From Enterprise Models to Dimensional Models: A Methodology for Datawarehouses and Data Mart Design. Proc. of the Int'l Workshop on Design and Management of Datawarehouses, Stockholm, Sweden, 2000, pp. 5.1-5.12.

[Pereira 2000] W. A. L. Pereira: A Methodology Targeted at the Insertion of Datawarehouse Technology in Corporations. MSc. Dissertation. Porto Alegre-PUCRS, 2000.

[Phipps 2002] C. Phipps, K. C. Davis: Automating Datawarehouse Conceptual Schema Design and Evaluation – Caise 2002.

[Poe 1998] V. Poe, P. Klauer, S. Brobst: Building a datawarehouse for decision support. New Jersey, Prentice Hall PTR, 1998.

[Rizzi 2002] S. Rizzi, M. Golfarelli, E. Saltarelli: Index selection for data warehousing – Caise 2002.

[Sapia 1998] C. Sapia, M. H. Blaschka, G. Fling, and B. Dinter: Extending the E/R Model for the Multidimensional Paradigm. Proc. of the Int'l Workshop on Data Warehousing and Datamining, Singapore, 1998, pp. 105-116.

[Semann 2000] H. B. Semann, J. Lechtenberger, and G. Vossen: Conceptual Datawarehouse Design. Proc. of the Int'l Workshop on Design and Management of Datawarehouses, Stockholm, Sweden, 2000, pp. 6.1-6.11.

[Song 2001] I.-Y Song, W.Rowen, C.Medsker, and E. Ewen: An Analysis of Many-to- Many Relationships Between Fact and Dimension Tables in Dimension Modeling. Proc. of the Int'l Workshop on Design and Management of Datawarehouses, Interlaken, Switzerland, 2001, pp. 6.1-6.13.

[Trujillo 2001] J.Trujillo, M.Palomar, J. Gomez, and I.-Y. Song: Designing Datawarehouses with OO Conceptual Models. IEEE Computer 34, 12 (2001), 66-75.

[Tsois 2002] A. Tsois, N. Karayannidis, T. Sellis, D. Theodoratos: Cost-Based Optimization of Aggregation Star Queries on Hierarchically Clustered Datawarehouses – Caise 2002